Statistically Efficient Estimation for Non-Smooth Probability Densities

Masaaki Imaizumi
Institute of Statistical Mathematics
RIKEN Advanced Intelligence Project

Takanori Maehara
RIKEN Advanced Intelligence Project

Yuichi Yoshida
National Institute of Informatics

Abstract

We investigate statistical efficiency of estimators for non-smooth density functions. The density estimation problem appears in various situations, and it is intensively used in statistics and machine learning. The statistical efficiencies of estimators, i.e., their convergence rates, play a central role in advanced statistical analysis. Although estimators and their convergence rates for smooth density functions are well investigated in the literature, those for non-smooth density functions remain elusive despite their importance in application fields. In this paper, we propose new estimators for non-smooth density functions by employing the notion of *Szemerédi partitions* from graph theory. We derive convergence rates of the proposed estimators. One of them has the optimal convergence rate in minimax sense, and the other has slightly worse convergence rate but runs in polynomial time. Experimental results support the theoretical performance of our estimators.

1 Introduction

Density estimation is one of the most significant and fundamental topics in statistical science and machine learning. Suppose that we have independent and identically distributed n observations:

$$X_1, X_2, \dots, X_n \sim F,$$

where F is a probability distribution function, and the probability density function f is defined by F. The goal of density estimation is to estimate f from the observations $\{X_i\}_{i=1}^n$. This density estimation problem is employed in many application fields such as obtaining spectrum densities in signal processing, estimating the distribution of animals in zoology, and analyzing volatility in financial analysis. Numerous studies investigated its statistical properties (See [1, 10, 31, 32] for summaries). There are a plethora of methods for the density estimation problem, e.g., histogram estimators [3, 5, 12, 28], kernel density estimators [2, 8, 15, 36], and orthogonal-series estimators [7, 11, 39].

Statistical efficiency of an estimator plays a central role in statistical analysis. It denotes the convergence rate of the error incurred by the estimator, that is, the speed of the convergence of the error to zero as n increases. The notion of statistical efficiency is used in various advanced inferences such as tests, confidence analysis, hyperparameter tuning, and others (summarized in [37, 40]). Clarifying the statistical efficiencies of estimators is essential to control uncertainty.

The *smoothness* of a density function has been an essential factor to bound the statistical efficiency of an estimator. For example, when a density function on a D-dimensional space is β-times differentiable, the convergence rate of the kernel density estimator is $O(n^{-2\beta/(2\beta+D)})$ under some regularity conditions (see [37] for a summary). Similarly, statistical efficiencies of other estimators are clarified only when the density function is differentiable.

In contrast, when the density function is *non-smooth*, i.e., non-differentiable or discontinuous, no estimator has been developed with a provable convergence rate, though some consistency results are known [6, 26, 42]. We note that non-smooth density functions frequently appear in real data; for instance, the spectrum density function has sharp peaks for some materials, and the distribution of a financial data often has many discontinuous points.

In this work, we propose two new estimators for non-smooth density functions with provable statistical efficiencies. The first one is called the *minimized Szemerédi density estimator* (M-SDE) and the second one is called the *Voronoi Szemerédi density estimator* (V-SDE). These estimators approximate density functions using a histogram, where its cells are constructed from a *Szemerédi partition*, a notion developed in graph theory [13, 16, 20, 24, 25, 33, 34].

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84. Copyright 2018 by the author(s).

Because the existence of a Szemerédi partition does not rely on the smoothness of the density function, our estimators can be used for non-smooth density functions. Then, we derive convergence rates of our estimators, and show that the convergence rate of the M-SDE is optimal in the minimax sense. Furthermore, we discuss how to tune the histogram size and time complexities of our estimators. Our numerical experiments confirm the performance of our estimators.

The contributions of this work are summarized as follows.

- We propose two estimators called the M-SDE and the V-SDE. We derive their convergence rates on non-smooth density functions. This is a first work which clarifies statistical efficiency of estimators for non-smooth densities.
- For M-SDE, we show that its convergence rate is optimal in the minimax sense. We also provide a hyperparameter selection method that attains the optimal convergence rate. For V-SDE, we show that its convergence rate is slightly worse than that of M-SDE. However, it runs in a polynomial time in n.
- Experimental results show that the M-SDE successfully estimated non-smooth density functions. We also show that V-SDE has a comparable approximation performance with M-SDE.

2 Preliminaries

2.1 Notations

Let $\mathbb{R}_+ := \{x \in \mathbb{R} : x \geq 0\}$ be the set of nonnegative numbers, and $I := [0,1]$ be the unit interval. For a matrix A, $A_{i,j}$ denotes an (i,j)-th element of A. For a positive integer z, $[z] := \{1,2,\ldots,z\}$ is the set of positive integers no more than z. For a function $f : I \to \mathbb{R}$, $\|f\|_1 := \int_I |f(t)|dt$ denotes the L^1 norm. λ denotes the Lebesgue measure. For a set S, $\mathbb{1}_{\{S\}}(x)$ is an indicator function which is 1 if $x \in S$ and 0 otherwise. For an element $s \in S$, a subset $S' \subseteq S$, and a metric $d : S \times S \to \mathbb{R}_+$, we define $d(s, S') := \min_{s' \in S'} d(s, s')$.

We recall Landau notations. For functions $f, g : \mathbb{N} \to \mathbb{R}_+$, $f(n) = \Omega(g(n))$ means $g(n) = O(f(n))$ and $f(n) = \Theta(g(n))$ means $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$. $\widetilde{O}(\cdot)$ and $\widetilde{\Theta}(\cdot)$ ignore the $O(\log\log n)$-factor.

2.2 Density Estimation Problem

We formulate the density estimation problem. Suppose that we observe n independent and identically distributed D-dimensional random variables:

$$X_1, X_2, \ldots, X_n \sim F^* \text{ on } (I^D, \mathcal{A}),$$

where F^* is an unknown true probability distribution, and $(I^D, \mathcal{A})$ is a measurable space with a compact set I^D and its σ-algebra $\mathcal{A}$. The true density function $f^* : I^D \to \mathbb{R}_+$ is defined using the Radon-Nikodym derivative of F^* with respect to the Lebesgue measure λ as

$$f^* = \frac{dF^*}{d\lambda}.$$

Note that f^* satisfies $\int_{I^D} f^* d\lambda = 1$.

The goal of the density estimation problem is to estimate f^* from the set of observations $\{X_i\}_{i\in[n]}$. The methodological and theoretical aspects of this problem have been investigated in the literature [1, 10, 31, 32].

2.3 Szemerédi Partitions

We introduce the notion of a Szemerédi partition that enables us to approximate a broad class of functions. The idea comes from the problem of approximating a graph by its small induced subgraph.

For partitions $\mathcal{P}_1, \ldots, \mathcal{P}_D$ of I, we define a partition $\mathcal{S}(\mathcal{P}_1, \ldots, \mathcal{P}_D)$ of I^D as

$$\mathcal{S}(\mathcal{P}_1, \ldots, \mathcal{P}_D) := \mathcal{P}_1 \times \cdots \times \mathcal{P}_D.$$

Each element in the partition $\mathcal{S} = \mathcal{S}(\mathcal{P}_1, \ldots, \mathcal{P}_D)$ is called a *cell*. Let $f_{\mathcal{S}}$ denote the *step function on* $\mathcal{S}$ of a function $f : I^D \to \mathcal{R}$ defined as

$$f_{\mathcal{S}}(x) := \sum_{S\in\mathcal{S};\lambda(S)>0} \frac{\mathbb{1}_S(x)}{\lambda(S)} \int_S f(y)dy.$$

Roughly speaking, $f_{\mathcal{S}}$ is obtained from f by taking the average in each cell $S \in \mathcal{S}$.

For partitions $\mathcal{P}_1, \ldots, \mathcal{P}_D$ of I, let $\mathcal{S} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_D$. For $\alpha > 0$, we say that a partition $\mathcal{S}$ is an *α-Szemerédi partition* of I^D with respect to f if

$$\sup_{\{T_d \subseteq I\}_{d\in[D]}} \left| \int_{T_1\times\cdots\times T_D} (f - f_{\mathcal{S}}) d\lambda \right| \leq \alpha. \qquad (1)$$

and that cells in $\mathcal{S}$ are *α-Szemerédi cells*. Here, T_d is taken from all measurable subsets of I for $d \in [D]$.

By the notion of Szemerédi partitions, we state an important result about the partitions.

Lemma. *(The Regularity Lemma* [24, 25]*) For any measurable bounded function f on I^D, there exist (equi)partitions $\mathcal{P}_1, \ldots, \mathcal{P}_D$ of I into K parts such that the partition $\mathcal{S} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_D$ of I^D is an $O(1/\sqrt{D \log K})$-Szemerédi partition with respect to f.*

3 Szemerédi Density Estimators

We explain our density estimation methods based on Szemerédi partitions.

Given observations $X_1, \dots, X_n$ and a (Szemerédi) partition $\mathcal{S}$ of I^D, our methods construct a *histogram function* $h_{\mathcal{S}} : I^D \to \mathbb{R}_+$ defined as

$$h_{\mathcal{S}}(x) := \frac{1}{n} \frac{\sum_{S \in \mathcal{S}} \mathbb{1}_S(x) \sum_{i \in [n]} \mathbb{1}_{(X_i \in S)}}{\sum_{S \in \mathcal{S}} \mathbb{1}_S(x) \lambda(S)}. \qquad (2)$$

This histogram counts the number of observations in each cell $S \in \mathcal{S}$ and normalizes it according to its size (the Lebesgue measure of S).

In what follows, we propose two estimators, both of which finds a partition $\mathcal{S}$ and returns the histogram function $h_{\mathcal{S}}$.

3.1 Minimized Szemerédi Density Estimator (M-SDE)

Here, we explain our first estimator, named the *minimized Szemerédi density estimator* (M-SDE). The M-SDE constructs a Szemerédi partition by combining sets in a finer partition of I. This method is standard in graphon analysis (see [24] for more details).

Let $K \in \mathbb{N}$ be a number of cells in $\mathcal{P}_d$ for each $d \in [D]$. For preparation, we consider a random partition with $M = zK$ cells, where $z \geq 2$ is a positive integer. Let $r_1, \dots, r_M$ be random variables uniformly sampled from I. Using these random variables, we define a random interval

$$R_m = \{x \in I : |x - r_m| \leq |x - r_{m'}|, \forall m' \in [M]\},$$

for each $m \in [M]$, and define $\mathcal{R} := \{R_m\}_{m \in [M]}$. As the Lebesgue measure of each R_m is roughly $1/M$ when M is large, we can roughly express a set with Lebesgue measure $1/K$ by combining M/K sets in $\mathcal{R}$. This fact suggests using an equipartition of $\{R_m\}_{m \in [M]}$ into K parts to roughly represent an equipartition of I into K parts. Let $\mathcal{P}_d^{(R)}$ be a partition of I which is generated from a K equipartition of $\{R_m\}_{m \in [M]}$.

To obtain partitions $\mathcal{P}_1, \dots, \mathcal{P}_D$ such that $\mathcal{S}(\mathcal{P}_1, \dots, \mathcal{P}_D)$ is an α-Szemerédi partition for small $\alpha > 0$, we employ the κ-fold cross-validation method. We split the observations into subsamples and generate κ pairs of a training data $\mathcal{D}^t$ and a validation data $\mathcal{D}^v$, denoted by $\{\mathcal{D}_j^t, \mathcal{D}_j^v\}_{j \in [\kappa]}$. Also, we define $\mathcal{S}^{\mathcal{R}} := \mathcal{S}(\mathcal{R}, \dots, \mathcal{R})$ as the partition of I^D constructed from $\mathcal{R}$. We define the validation function on partitions of I^D with respect to the j-th pair as

$$\mathrm{CV}_j(\mathcal{S}) := \int_{I^D} \left| h_{\mathcal{S}^{\mathcal{R}}, \mathcal{D}_j^v}(x) - h_{\mathcal{S}, \mathcal{D}_j^t}(x) \right| dx, \qquad (3)$$

where $h_{\mathcal{S}, \mathcal{D}}$ is the histogram function (2) with the partition $\mathcal{S}$ and the data $\mathcal{D}$. Then, we define the cross-validation problem as

$$\min_{\mathcal{P}_1^{(R)}, \dots, \mathcal{P}_D^{(R)}} \frac{1}{\kappa} \sum_{j \in [\kappa]} \mathrm{CV}_j \left(\mathcal{S}(\mathcal{P}_1^{(R)}, \dots, \mathcal{P}_D^{(R)}) \right). \qquad (4)$$

This optimization problem (4) is an approximated version of $\min_{\mathcal{S}} \|f^* - f_{\mathcal{S}}\|_1$, i.e., we approximate $\mathcal{S}$ by $\mathcal{S}(\mathcal{P}_1^{(R)}, \dots, \mathcal{P}_D^{(R)})$ and f^* by the empirical distribution. Theorem 1 in Section 4 will provide its validity.

We set a minimizer of (4) as $\mathcal{S}^M$, and define the minimized Szemerédi density estimator (M-SDE) as

$$\widehat{f}^M(x) := h_{\mathcal{S}^M}(x). \qquad (5)$$

In Section 4, we will provide a methodology for selecting hyper-parameters and will show theoretical properties of the methodology.

We mention that there are some limitations of M-SDE. Solving the minimization problem (4) requires an exponential time in n and a global convergence property is not assured. To avoid the difficulties, we will suggest another estimator in the next section.

3.2 Voronoi Szemerédi Density Estimator (V-SDE)

We provide another estimator named the Voronoi Szemerédi density estimator (V-SDE). It has slightly worse convergence rate than M-SDE but can be constructed in polynomial time in n.

The V-SDE is based on Voronoi partitions of $\mathcal{R}$ which is introduced in Section 3.1 . For the V-SDE, we consider the case $D = 2$. Also, we assume that f^* is symmetric around, that is, $f^*(x, y) = f^*(y, x)$ in this section. We note that, to estimate asymmetric density functions, we can repeat our estimation method on $\{(x, y) \in I^2 : x \geq y\}$ and $\{(x, y) \in I^2 : x < y\}$ separately.

We introduce the Voronoi partition induced by a set of indices $\mathcal{V} \subseteq [M]$. First, we define a metric $d_A : [M] \times [M] \to \mathbb{R}_+$ as

$$d_A(a, b) = \sum_{i \in [M]} \left| \sum_{j \in [M]} A_{j,i}(A_{j,a} - A_{j,b}) \right|$$

where A is an $M \times M$ matrix with $A_{i,j} = \int_{R_i} \int_{R_j} f^*(x, y) dy dx$. Then, we define $c_{A,\mathcal{V}} : [M] \to [M]$ so that $c_{A,\mathcal{V}}(m)$ is the index in $\mathcal{V}$ closest to m with respect to d_A, where we break ties arbitrarily. The *Voronoi partition* of I generated by $\mathcal{R}$, $\mathcal{V}$, and d_A is defined as

$$\mathcal{P}(\mathcal{V}) := \left\{ \bigcup_{m \in N_\delta(\ell)} R_m : \ell \in \mathcal{V} \right\},$$

where $N_\delta(\ell) := \{m \in [M] : c_{A,\mathcal{V}}(m) = \ell\}$ is the Voronoi cell centered at $\ell \in \mathcal{V}$.

We can show that $\mathcal{S}(\mathcal{P}(\mathcal{V}), \mathcal{P}(\mathcal{V}))$ is a Szemerédi partition with an existing $\mathcal{V}$ by using results in [24]. Specifically, for

Algorithm 1 Generating indices for Voronoi partition

```
input  $\epsilon_n \in (0,1)$, $\delta \in (0,1)$ and $d_{\widehat{A}}$.
output  A set $\mathcal{V} \subseteq [M]$.
$\mathcal{V} \leftarrow \emptyset$.
$C_T \leftarrow \epsilon_n^{-3}\log(1/\epsilon_n) + \epsilon_n^{-1}\log(1/\delta)$.
$c \leftarrow 0$.
while $c \leq C_T$ do
    $\ell \leftarrow$ an index uniformly sampled from $[M]$.
    if $d_{\widehat{A}}(\ell, \mathcal{V}) > \epsilon_n/2$ or $d_{\widehat{A}^\top}(\ell, \mathcal{V}) > \epsilon_n/2$ then
        $\mathcal{V} \leftarrow \mathcal{V} \cup \{\ell\}$.
    else
        $c \leftarrow c + 1$.
    end if
end while
```

any $\epsilon > 0$, there exists $\eta > 0$ such that, if $\mathcal{V}$ satisfies

$$\sum_{m\in[M]} \min_{\ell\in\mathcal{V}} d_A(m,\ell) < \eta, \tag{6}$$

then we have $\|f^* - f^*_{\mathcal{S}(\mathcal{P}(\mathcal{V}),\mathcal{P}(\mathcal{V}))}\|_1 \leq \epsilon$.

To estimate $\mathcal{V}$ satisfying (6), we provide an algorithm with an empirical metric. As we cannot exactly calculate elements of the matrix A, we consider an empirical metric $d_{\widehat{A}} : [M] \times [M] \to \mathbb{R}_+$ defined as

$$d_{\widehat{A}}(a,b) := \sum_{i\in[M]} \left| \sum_{j\in[M]} \widehat{A}_{j,i}(\widehat{A}_{j,a} - \widehat{A}_{j,b}) \right|.$$

Here, $\widehat{A}$ is an $M \times M$ matrix generated from the observations:

$$\widehat{A}_{i,j} = (n\lambda(R_i \times R_j))^{-1} \sum_{i=1}^{n} \mathbb{1}_{(X_i \in R_i \times R_j)}.$$

Algorithm 1 provides a pseudo-code for constructing $\mathcal{V}$ satisfying (6) using $d_{\widehat{A}}$ and predetermined parameters $\epsilon_n \in (0,1)$ and $\delta \in (0,1)$ for controlling convergence of V-SDE. Also, we define C_T in Algorithm 1 as a maximum iteration defined by ϵ_n and δ. Since $\widehat{A}$ converges to A as n increases, we can bound the difference between d_A and $d_{\widehat{A}}$ for sufficiently large n. The analysis of this difference is deferred to Section 4.

Let $\widehat{\mathcal{V}}$ be an output of Algorithm 1, the Voronoi Szemerédi density estimator (V-SDE) is defined as

$$\widehat{f}^V(x) := h_{\mathcal{S}(\mathcal{P}(\widehat{\mathcal{V}}),\mathcal{P}(\widehat{\mathcal{V}}))}(x). \tag{7}$$

The time complexity of the V-SDE is polynomial in n: Constructing $\mathcal{V}$ requires $O(n + M^2)$ preprocessing and $O((C_T + M)M^3)$ time, and constructing $\mathcal{P}(\mathcal{V})$ requires $O((C_T + M)M^3)$ time. In Section 4, we will show that it suffices to set $M = \Theta(K) = \Theta\left(n^{1/2D}\right)$. Also, we will set $\epsilon_n = O\left((\log n)^{-1/4}\right)$ and thus $C_T = O(\log n)$. Hence, the time complexity is polynomial in n.

4 Convergence Analysis

In this section, we analyze convergence rates of the M-SDE and the V-SDE. Moreover, for the M-SDE, we also show its minimax optimality and the effect of hyperparameter selection. All the proofs are deferred to the supplementary material.

We evaluate the convergence rate of the estimator $\widehat{f} \in \{\widehat{f}^M, \widehat{f}^V\}$ by the L^1-loss

$$\|f^* - \widehat{f}\|_1.$$

This is because the L^1-loss is related to the *total variation* distance for probability distribution functions. Let F and F' be probability distribution functions on a measurable space $(I^D, \mathcal{A})$. Then, the total variation distance between F and F' is defined as

$$\|F - F'\|_{\mathrm{TV}} := \sup_{A\in\mathcal{A}} |F(A) - F'(A)|.$$

Furthermore, let f and f' be the density functions obtained through the Radon-Nikodym derivatives of F and F', respectively. Then, the total variation distance is equivalently written as

$$\|F - F'\|_{\mathrm{TV}} = \frac{1}{2}\int_{I^D} |f(x) - f'(x)|dx = \frac{1}{2}\|f - f'\|_1$$

by using Scheffe's theorem (Lemma 2.1 in [37]).

We impose the following two conditions on the true density function f^*.

Assumption 1. *The following conditions hold.*

1. *The density function is defined as the Radon-Nikodym derivative, that is, $f^* = dF^*/d\lambda$, where F^* is a probability measure on the measurable space $(I^D, \mathcal{A})$.*
2. *f^* is a bounded function on I^D.*

The first condition is a standard regularity condition on probability density functions. We note that the second boundedness condition is quite weaker than other conditions such as continuity and differentiability.

4.1 Convergence Rate of the M-SDE

In this section, we evaluate the convergence rate of the M-SDE.

First, we show that, with high probability, we can construct a Szemerédi partition from the set of random intervals $\mathcal{R} = \{R_m\}_{m\in[M]}$ introduced in Section 3.1:

Lemma 1. *Suppose Assumption 1 holds. Let $\mathcal{R} = \{R_m\}_{m\in[M]}$ be the set of random intervals introduced in Section 3.1. Then with probability $1 - O(1/M^2)$, there exists a set of partitions $\{\mathcal{P}_d\}_{d\in[D]}$ obtained by combining* *sets in $\mathcal{R}$ such that the partition $\mathcal{S} = \mathcal{S}(\mathcal{M}_1, \dots, \mathcal{M}_D)$ of I^D satisfies*

$$\sup_{\{T_d \subseteq I\}_{d\in[D]}} \left| \int_{T_1\times\cdots\times T_D} (f - f_{\mathcal{S}}) d\lambda \right| = O\left(\frac{1}{\sqrt{D \log M}} \right).$$

The proof of Lemma 1 is based on the sampling lemma for graphons (see Section 10 in [24]). As we set $M = zK$ with a positive integer z, the approximation error is bounded by $O(1/\sqrt{D\log K})$.

Then, we have the following convergence rate.

Theorem 1. *Consider the estimator* (5). *Then for any $K \geq 1$ and sufficiently large n, we have*

$$\|\widehat{f}^M - f^*\|_1 = O\left(\frac{K^D \log n}{\sqrt{n}} + \frac{1}{\sqrt{D \log K}} \right),$$

with probability at least $1 - O(1/(nM)^2)$.

The first term of this bound is the estimation error. In particular, the first term bounds the distance $\|\widehat{f}^M - f^*_{\mathcal{S}}\|_1$ where $f^*_{\mathcal{S}}$ is a step function generated from the true density f^*. As the estimation error is measured on each cell $S \in \mathcal{S}$, the error is increased with the number of cells. When K is constant, the estimation error attains the parametric rate $O(n^{-1/2})$.

The second term represents the approximation error $\|f^*_{\mathcal{S}} - f^*\|_1$. This error is evaluated through the weak regularity lemma (see, e.g., [24]) and is bounded by $O(1/\sqrt{\log K})$.

Now we take K as a function of n to balance the estimation and approximation errors. Specifically, we consider an increasing sequence $\{K_n\}_{n\in\mathbb{Z}}$ satisfying the following condition:

$$\frac{K_n^D \log n}{\sqrt{n}} = \Theta\left(\frac{1}{\sqrt{D \log K_n}} \right). \tag{8}$$

Then we have $K_n = \widetilde{\Theta}(n^{1/2D}(\log n)^{-3/2D})$. By substituting this to Theorem 1, we obtain the following corollary.

Corollary 1. *For each n, by choosing $K = K_n = \Theta(n^{1/2D}(\log n)^{-3/2D})$ in the M-SDE, the estimator* (5) *satisfies*

$$\|\widehat{f}^M - f^*\|_1 = \widetilde{O}\left(\frac{1}{\sqrt{\log n}} \right),$$

with probability larger than $1 - O(1/(nM)^2)$.

Corollary 1 states that the statistical efficiency of the M-SDE is $O\left(1/\sqrt{\log n}\right)$, which is independent of D. Although this speed of convergence is slow, we will discuss its optimality and rationale in Sections 4.4 and 4.5.

4.2 Selection of the Partition Size

We propose a method that selects a suitable K to achieve the convergence rate in Corollary 1 by exploiting the condition (8) that K should satisfy.

We use the decomposition of the L^1-loss of the M-SDE as seen in Theorem 1. The loss is decomposed into the estimation error and the approximation error. Hence, we evaluate those errors empirically and choose K to balance them. Our approach employs Lepski's method [23] and can achieve the optimal convergence rate as in Corollary 1.

For clarity of the discussion, we let $\widehat{f}^{M,(K)}$ be the M-SDE with a partition of size K. Let $K_{\max}$ be a large integer. Then, the estimator is defined as

$$\widehat{K}_n = \min\Big\{ K \in [K_{\max}] \mid \forall \ell > K, \ell \in [K_{\max}], \|\widehat{f}^{M,(K)} - \widehat{f}^{M,(\ell)}\|_1 \leq \tau n^{-1/2} \ell^D \log n \Big\}, \tag{9}$$

where $\tau > 4$ is an arbitrary constant. The threshold term, i.e., $\tau n^{-1/2}\ell^D \log n$ in (9), is derived from the condition (8). If the set in (9) is empty, then we set $\widehat{K}_n = K_{\max}$.

The following theorem shows that the M-SDE with the selection method above can achieve the minimax optimal rate derived in Corollary 1.

Theorem 2. *Suppose that Assumption 1 holds and we choose the size of the partition as in* (9). *Then, we have*

$$\|\widehat{f}^{M,(\widehat{K}_n)} - f^*\|_1 = \widetilde{O}\left(\frac{1}{\sqrt{\log n}} \right),$$

with probability $1 - O(1/(nM)^2)$ for sufficiently large n.

Theorem 2 shows that we can achieve the convergence rate in Corollary 1.

4.3 Convergence Rate of the V-SDE

Next, we evaluate the convergence of the V-SDE. Applying Lemma 1, we obtain the following.

Theorem 3. *Consider the estimator* (7) *with $K_n = \Theta(n^{1/2D}(\log n)^{-3/2D})$ and $\epsilon_n = \delta(\log n)^{-1/4}$ for $\delta \in (0,1)$. Suppose Assumption 1 holds. Then, there exists $C_V = C_V(D) > 0$ such that*

$$\|\widehat{f}^V - f^*\|_1 = O\left(\frac{C_V}{(\log n)^{1/8}} \right),$$

holds with probability $1 - O(\delta/(nM)^2)$ for sufficiently large n.

The convergence rate provided in Theorem 3 is slower than that of Corollary 1, in exchange for computational efficiency of the Voronoi approximation.

4.4 Minimax Optimality

We investigate a minimax lower bound on the density estimation problem. To this end, we consider a proper subclass of density functions and derive a lower bound on the convergence rate of the estimator for this subclass.

We introduce a coefficient $g_S \in \mathbb{R}_+$ and a binary variable θ_S for each cell S in $\mathcal{S}$. Let

$$\widetilde{\mathcal{F}} := \Bigg\{ f : I^D \to \mathbb{R}_+ \mid f(x) = \sum_{S \in \mathcal{S}} \mathbb{1}_{(x \in S)} g_S \theta_S, \\ \exists S' \in \mathcal{S}, g_{S'} = \frac{C_O}{\sqrt{D \log K_n}} \Bigg\},$$

for some constant $C_O > 0$. Intuitively, each density function in $\widetilde{\mathcal{F}}$ is a step function with respect to $\mathcal{S}$ with each cell $S \in \mathcal{S}$ having the value g_S except $g_{S'}$ for some $S' \in \mathcal{S}$ having a specific value.

The following theorem provides the minimax lower bound for density estimators.

Theorem 4. *Suppose that Assumption 1 holds. Then, we have*

$$\liminf_{n \to \infty} \inf_{\overline{f}_S} \sup_{f \in \widetilde{\mathcal{F}}} \mathrm{P}\left(\|\overline{f}_{\mathcal{S}} - f\|_1 \geq \frac{C_O}{\sqrt{D \log K_n}} \right) > 0,$$

where $\overline{f}_{\mathcal{S}}$ runs over step function estimators with respect to the partition $\mathcal{S}$.

Substituting $K_n = \Theta(n^{1/2D}(\log n)^{-3/2D})$ yields that the lower rate in Theorem 4 is $\widetilde{\Theta}(1/\sqrt{\log n})$. This result shows that the convergence rates in Corollary 1 is minimax optimal.

4.5 Discussions

The convergence rate of the M-SDE attains the minimax optimality, and also we can conduct the parameter selection preserving the convergence rate. One can criticize that the convergence rate of $O\left((\log n)^{-1/2}\right)$ is slower than those of estimators for smooth densities, which are polynomial in n. However, recalling that the class of non-smooth density functions is quite larger than that of smooth ones, it is natural to have an exponentially weaker rate. We also mention a study for estimating non-smooth functional [4], which clarifies that the optimal convergence rate is $O\left((\log n)^{-1/2}\right)$.

The V-SDE yields the convergence rate of $O((\log n)^{-1/8})$, which is $O((\log n)^{3/8})$ times slower than the M-SDE. However, the time complexity of the V-SDE is polynomial in n, which is exponentially faster than that of the M-SDE.

5 Comparison with Related Work

We clarify the differences between our density estimators and other density estimators.

Various studies have investigated histogram-based density estimators. The simplest one exploits the histogram with fixed cells generated from a partition $\{[t_j, t_{j+1})\}_{j \in [K]}$ in to K cells, where $t_j < t_{j+1}$ for all $j \in [K]$. The consistency of this estimator is shown in several situations: [6] and [26] showed the consistency with the L^1-norm and [42] showed the consistency with the L^p-norm. The convergence speed is also a concern for this histogram estimator. By assuming the differentiability of density functions with a scalar input setting, [9] and [12] showed that the convergence rate for the differentiable density functions is $O(n^{-1/3})$. Some studies [29, 30] showed the convergence rate can be improved to $O(n^{-2/5})$ by modifying the arrangement of the cells. However, the convergence rate *without* the differentiability assumption remains elusive.

Other studies investigated a histogram estimator with data-adaptive cells, known as a *variable partition histogram* [41]. This estimator constructs cells from the observations. For example, let $\{X_{(1)}, X_{(2)}, \ldots, X_{(n)}\}$ be a set of ordered observations (i.e., $X_{(i)} \leq X_{(i+1)}$ for all i), then the j-th cell (out of K cells) is defined as $(X_{(jn/K)}, X_{(jn/K)}]$ for $j \in [K]$. [17, 19, 21, 27] showed the consistency of this estimator with respect to various norms under various conditions. Moreover, assuming the differentiability of density functions, the convergence rate of this estimator was clarified in [18]. However, as with the estimator based on a histogram with fixed cells, the convergence rate is not obtained without the differentiability assumption.

Recently, density estimation methods using functional components have been intensively studied, e.g., the kernel density estimator, orthogonal series estimator, and Gaussian process estimator. The *kernel density estimator* [2, 8, 15, 36] estimates density functions by using a kernel function. It is shown that the kernel density estimator has a convergence rate of $O(n^{-1/3})$ when the true density is differentiable and the kernel function is properly selected. Similar results were obtained for the orthogonal series estimator [7, 11, 39] and the Gaussian process estimator [22, 38]. We stress again that these methods assume that the true density is differentiable several times. In contrast, SDEs only assume that the density function is measurable and bounded.

The aforementioned comparison is summarized in Table 1.

6 Experiments

We conduct numerical experiments to validate the theoretical results on SDEs.

6.1 Density Estimation

The goal of this section is to confirm that our methods well approximate density functions. To show this, we generated $10,000$ observations from three types of density functions; (A) the Gaussian type, (b) the cylinder type, and (C) the pyramid type, and estimated these density functions by using the M-SDE. The size K of a partition was selected so as to minimize the generalization error, and we set the number

Estimator	Differentiable Density	Continuous, but Non-Differentiable Density	Discontinuous Density
Histogram (fixed cell)	√	(Consistency)	(Consistency)
Histogram (adaptive cell)	√		
Kernel	√		
Orthonormal Series	√		
Gaussian Process	√		
Szemerédi (M-SDE & V-SDE)	√	√	√

Table 1: Statistical efficiency of each estimator on each class of density functions. "√" means that a statistical efficiency can be derived. All estimators have provable statistical efficiencies for differentiable densities whereas only SDEs have provable statistical efficiencies for non-differentiable or discontinuous densities.

of random variables M as $M = 3K$. To solve the minimization problem (4), we employed a greedy algorithm by swapping. This algorithm starts with an arbitrary partition and then keep swapping two intervals in different sets in the partition until the objective function converges.

For each setting, we plot the true density function and its approximation by the M-SDE. For the cell plots, blue and denote small and large values, respectively.

Gaussian-type Density We consider a Gaussian-type density function on $[0,1]^2$ visualized in Figure 1. This density function is a two-variate Gaussian restricted to $[0,1]^2$. The selected size of the partition was $K = 4$, which is relatively small due to the smoothness of the Gaussian-type density. As can be seen in Figure 1, the cells can capture the shape of the Gaussian-type density, that is, the cells close to the edges of $[0,1]^2$ are narrow whereas the cells around the center of $[0,1]^2$ are close to squares.

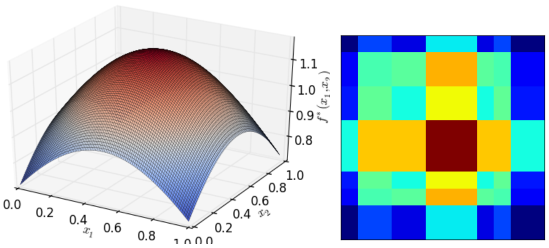

Figure 1: Gaussian-type density function (left), and its approximation by the M-SDE (right). K was set to 4.

Cylinder-type Density We consider a cylinder-type density function on $[0,1]^2$ visualized in Figure 2. The density function takes a positive value inside a circle in $[0,1]^2$ and takes zero elsewhere. Its approximation by the M-SDE is also shown in Figure 2. The selected size of the partition was $K = 10$ due to the discontinuous structure of the density function. We also can see that the cells constructed by the algorithm express the round shape of the true density.

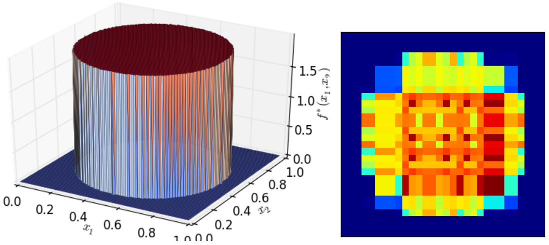

Figure 2: Cylinder-type density function (left), and its approximation by the M-SDE (right). K was set to 10.

Pyramid-type Density We consider a pyramid-type density function on $[0,1]^2$ visualized in Figure 3. The density function is constituted by four sharp pyramids. Its approximation by the M-SDE is also shown in Figure 3. Since the pyramid-type density is highly non-smooth, the selected size of the partition was $K = 13$, which is larger than the other two cases.

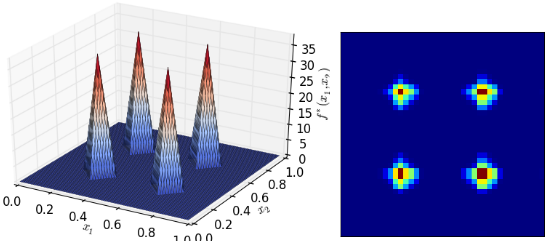

Figure 3: Pyramid-type density function (left), and its approximation by the M-SDE (right). K was set to 13.

6.2 Comparison with Other Estimators

We compare the performance of our estimators with that of other estimators. For each $n \in \{1000, 2000, \ldots, 10000\}$, we generated n independent observations from a true density function f^*, which is one of the three density functions used above, and estimated f^* with the following density estimators. All hyperparameters were selected to minimize the generalization error.

Kernel density estimator: We employed the Gaussian kernel $k_h(x) \propto \exp(-x^2/2h^2)$ (Kernel(gauss)) and the square kernel $k_h(x) \propto \mathbb{1}_{(x<h)}$ (Kernel(square)) for estimation. The bandwidth h of the kernel functions was selected from $\{0.001, 0.005, 0.01, 0.02, \ldots, 0.1\}$.

Histogram estimator (Histogram) with equal-sized cells: We used the partition $\{[0, 1/K), [1/K, 2/K), \ldots [(K-1)/K, 1]\}^2$, where $K \geq 2$ is an integer, resulting in K^2 cells. K was selected from $\{1, 2, \ldots, 20\}$. We do not consider the variable partition histogram method since it can be simulated by Kernel(square).

Orthogonal Series density estimator (Series): We employed the Fourier basis function for this estimator. This estimator represents a density function as $f(x) = \sum_j w_j \phi_j(x)$, where $\phi_j(x)$ is the Fourier basis functions and w_j is a weight for each j. The estimator computes the weight w_j by loss minimization. The number of basis functions was selected from $\{5, \ldots, 20\}$.

Figure 4 shows generalization errors as expected L^1 losses of the estimators on each density function. For the Gaussian-type case, the kernel density estimators showed significantly large losses, and other methods can estimate the smooth Gaussian function accurately. For the cylinder-type case, all the methods showed similar performances. For the pyramid-type case, the M-SDE, the V-SDE, and the kernel method with the square kernel can well estimate the sharp poles in the density function.

Observations obtained from those results are in order. (i) For each case, the M-SDE and the V-SDE seem to be consistent. Some of the other estimators seem biased, especially when estimating the pyramid-type density. In contrast, the losses of the M-SDE and the V-SDE always decrease as n increases. (ii) The numerical convergence speeds of the estimators are similar although only the M-SDE and the V-SDE have theoretical guarantees. (iii) The V-SDE has a similar numerical performance to the M-SDE although the theoretical efficiency of the V-SDE is worse than that of the M-SDE.

7 Conclusion

Estimation of density functions is a significant topic in statistics and machine learning, and it has been intensively investigated. Nevertheless, statistically efficient estimation of non-smooth density functions has been elusive. Filling this gap is important since statistical efficiency is a critical factor in advanced inferences such as tests or confidence analysis.

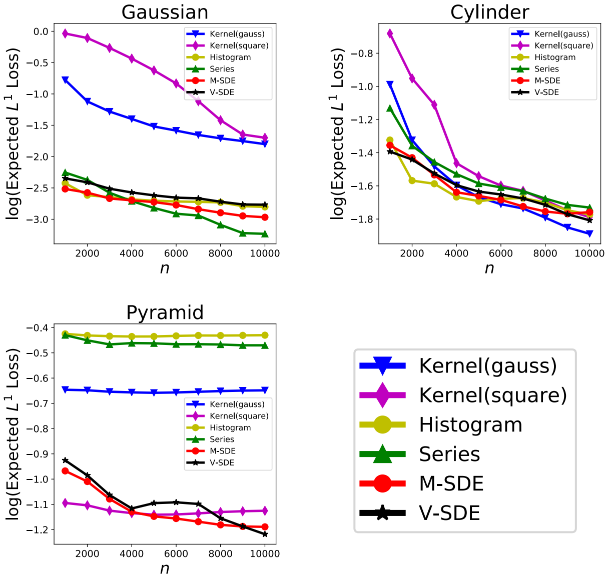

Figure 4: Logarithm of expected L^1 losses against n for the Gaussian case (upper left), the cylinder case (upper right), and the pyramid case (lower left).

To derive a statistically efficient estimation of non-smooth density functions, we propose the *minimized Szemerédi density estimator* (M-SDE) and the *Voronoi Szemerédi density estimator* (V-SDE). The convergence rate of the M-SDE attains the minimax optimal rate. We also proved that the selection method for the hyperparameter of the M-SDE does not deteriorate the convergence rate. Although the M-SDE is computationally expensive, the V-SDE which is computationally efficient at the cost of a relatively slow convergence rate. The experimental results validate the theoretical analysis of the M-SDE and the V-SDE.

Our work will facilitate statistical analysis on non-smooth density functions. Developing such statistical analysis is left for future work. A construction of a minimax optimal estimator with polynomial time in a sample size is an open question.

Acknowledgements

M. I. is supported by Grant-in-Aid for JSPS Research Fellow (17J03208) from the JSPS. T. M. is supported by JSPS KAKENHI Grant Number 16K16011. Y. Y. is supported by JSPS KAKENHI Grant Number JP17H04676 and JST ERATO Grant Number JPMJER1305, Japan.

References

[1] Richard E Barlow, David J Bartholomew, JM Bremner, and H Daniel Brunk, *Statistical inference under order restrictions: The theory and application of isotonic regression*, Wiley New York, 1972.

[2] Zdravko I Botev, Joseph F Grotowski, and Dirk P Kroese, *Kernel density estimation via diffusion*, The Annals of Statistics **38** (2010), no. 5, 2916–2957.

[3] T Bouezmarni, M Mesfioui, and JM Rolin, *L1-rate of convergence of smoothed histogram*, Statistics & probability letters **77** (2007), no. 14, 1497–1504.

[4] T Tony Cai and Mark G Low, *Testing composite hypotheses, hermite polynomials and optimal estimation of a nonsmooth functional*, The Annals of Statistics **39** (2011), no. 2, 1012–1041.

[5] Siu On Chan, Ilias Diakonikolas, Rocco A Servedio, and Xiaorui Sun, *Near-optimal density estimation in near-linear time using variable-width histograms*, Advances in neural information processing systems, 2014, pp. 1844–1852.

[6] XR Chen and LC Zhao, *Almost sure l 1-norm convergence for data-based histogram density estimates*, Journal of multivariate analysis **21** (1987), no. 1, 179–188.

[7] NN Chentsov, *Evaluation of an unknown distribution density from observations*, Doklady Akademii Nauk SSSR **147** (1962), no. 1, 45.

[8] Luc Devroye, *The equivalence of weak, strong and complete convergence in l_1 for kernel density estimates*, The Annals of Statistics **11** (1983), no. 3, 896–904.

[9] Luc Devroye and Laszlo Gyorfi, *Nonparametric density estimation: the l1 view*, Vol. 119, John Wiley & Sons Incorporated, 1985.

[10] Luc Devroye and Gábor Lugosi, *Combinatorial methods in density estimation*, Springer Science & Business Media, 2012.

[11] Sam Efromovich, *Adaptive estimation of and oracle inequalities for probability densities and characteristic functions*, The Annals of Statistics **36** (2008), no. 3, 1127–1155.

[12] David Freedman and Persi Diaconis, *On the maximum deviation between the histogram and the underlying density*, Probability Theory and Related Fields **58** (1981), no. 2, 139–167.

[13] Alan Frieze and Ravi Kannan, *A simple algorithm for constructing szemerédis regularity partition*, the electronic journal of combinatorics **6** (1999), no. R17, 2.

[14] James E Gentle, *Computational statistics*, Vol. 308, Springer, 2009.

[15] Alexander Goldenshluger and Oleg Lepski, *Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality*, The Annals of Statistics **39** (2011), no. 3, 1608–1632.

[16] W Timothy Gowers, *Hypergraph regularity and the multidimensional szemerédi theorem*, Annals of Mathematics **166** (2007), no. 3, 897–946.

[17] Yuichiro Kanazawa, *An optimal variable cell histogram*, Communications in Statistics—Theory and Methods **17** (1988), no. 5, 1401–1422.

[18] ______, *An optimal variable cell histogram based on the sample spacings*, The Annals of Statistics **20** (1992), no. 1, 291–304.

[19] Atsuyuki Kogure, *Asymptotically optimal cells for a historgram*, The Annals of Statistics **15** (1987), no. 3, 1023–1030.

[20] János Komlós and Miklós Simonovits, *Szemerédi's regularity lemma and its applications in graph theory* (1996).

[21] Jean-Pierre Lecoutre, ML Puri, JP Vilaplana, and W Wertz, *The histogram with random partition*, New Perspectives in Theoretical and Applied Statistics (1986), 265–276.

[22] Tom Leonard, *Density estimation, stochastic processes and prior information*, Journal of the Royal Statistical Society. Series B (Methodological) (1978), 113–146.

[23] Oleg V Lepskii, *On a problem of adaptive estimation in gaussian white noise*, Theory of Probability and its Applications **35** (1990), 454–466.

[24] László Lovász, *Large networks and graph limits*, Vol. 60, American Mathematical Society Providence, 2012.

[25] László Lovász and Balázs Szegedy, *Szemerédis lemma for the analyst*, GAFA Geometric And Functional Analysis **17** (2007), no. 1, 252–270.

[26] Gábor Lugosi and Andrew Nobel, *Consistency of data-driven histogram methods for density estimation and classification*, The Annals of Statistics **24** (1996), no. 2, 687–706.

[27] BLS Prakasa Rao, *Nonparametric functional estimation*, Academic press, 2014.

[28] David W Scott, *On optimal and data-based histograms*, Biometrika **66** (1979), no. 3, 605–610.

[29] ______, *Averaged shifted histograms: Effective nonparametric density estimators in several dimensions*, The Annals of Statistics **13** (1985), no. 3, 1024–1040.

[30] ______, *Frequency polygons: theory and application*, Journal of the American Statistical Association **80** (1985), no. 390, 348–354.

[31] ______, *Multivariate density estimation: theory, practice, and visualization*, John Wiley & Sons, 2015.

[32] Bernard W Silverman, *Density estimation for statistics and data analysis*, Vol. 26, CRC press, 1986.

[33] Endre Szemerédi, *Regular partitions of graphs*, DTIC Document, 1975.

[34] Terence Tao, *Szemer'edi's regularity lemma revisited*, Contributions to Discrete Mathematics **1** (2006), no. 1.

[35] ______, *Epsilon of room, two: pages from year three of a mathematical blog*, Vol. 1, American Mathematical Soc., 2011.

[36] George R Terrell and David W Scott, *Variable kernel density estimation*, The Annals of Statistics **20** (1992), no. 3, 1236–1265.

[37] Alexandre B Tsybakov, *Introduction to nonparametric estimation*, Springer Series in Statistics. Springer, New York, 2003.

[38] AW van der Vaart and JH van Zanten, *Rates of contraction of posterior distributions based on gaussian process priors*, The Annals of Statistics **36** (2008), no. 3, 1435–1463.

[39] Gilbert G Walter, *Properties of hermite series estimation of probability density*, The Annals of Statistics **5** (1977), no. 6, 1258–1264.

[40] Larry Wasserman, *All of nonparametric statistics*, 2005.

[41] Edward J Wegman, *Maximum likelihood estimation of a unimodal density function*, The Annals of Mathematical Statistics **41** (1970), no. 2, 457–471.

[42] Lin Cheng Zhao, Paruchuri R Krishnaiah, and Xi Ru Chen, *Almost sure l_r-norm convergence for data-based histogram density estimates*, Theory of Probability & Its Applications **35** (1991), no. 2, 396–403.